Visual Analytics Methods for the Automatic Content Generation from Streaming Data

Fabio Giachelle[0000−0001−5015−5498]

Department of Information Engineering, University of Padua
fabio.giachelle@unipd.it

Abstract. We present a PhD project regarding the application of Visual Analytics (VA) methods for the automatic generation of wiki documents - i.e. wikification - and event storylines from streaming data. In contrast to static automatically generated wiki-like documents, this project investigates the employment of VA techniques for the automatic generation of wiki documents made up of dynamic contents, based on user preferences. The purpose of the project is to make the user an active component for the wikification process, able to provide useful feedback regarding which contents are more relevant for the topic of interest, thus improving the wikification algorithms. For this purpose, the project focuses on exploiting VA methods and data provenance to enhance data comprehension, by means of continuous interaction with the user according to the human-in-the-loop model.

Keywords: Visual Analytics · Wikification · Data Provenance · Human-in-the-loop.

1 Introduction

1.1 Overall context

Nowadays, millions of users everyday surf the Internet looking for useful information to satisfy their information needs. In particular, Wikipedia is one of the most visited reference websites of all time and probably the most popular web-based, free-content encyclopedia of the world. Since Wikipedia is based on a model of openly editable content, the number of articles is growing continuously. In the last years, the automatic creation of Wikipedia articles – i.e. automatic wikification – has grown of interest. In particular, recent research works focus on the automatic creation of wiki documents, dynamically edited over time, based on distributed streaming data from heterogeneous sources as newsfeed and social media. However, these documents are available to end-users, only as static web pages. Unfortunately, in this way, users cannot provide any useful feedback to assess and improve the performances of the wikification algorithms. For this reason, as shown in Figure 1, the focus of this project is to allow the dynamic visualization of automatically generated articles, by means of interactive visual interfaces that control the algorithms underlying them. In this way, users can judge which contents are more relevant for the given topic and exploit implicit and explicit user feedback to assess and improve the performances of the wikification algorithms.

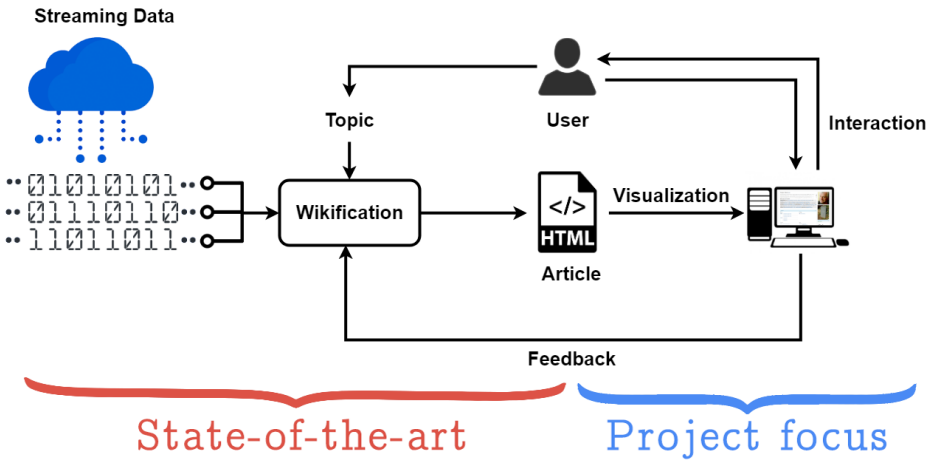

Fig. 1: Wikification and improved understanding of data by means of visual analytics methods.

1.2 State-of-the-art

Wikification and event storylines generation

The term "wikification" [12] refers to the creation of documents containing entities linked to Wikipedia, which represents the target knowledge base. The task of identifying entities in a given text and linking them to a specific knowledge base is known as "entity linking" [13]. Event storylines, instead, are a chronological reconstruction of a sequence of event happenings over time, related to a topic of interest [1]. In particular, we consider an "event" something that happens, important or of interest for users. Nowadays, news and information about events are shared mostly through social media. Hence, many research works focus on collecting useful information, from social network services, for the automatic generation of event storylines [10]. The main idea is to generate event storylines, by the fusion of crowdsourced retrieved data [3, 7] to grant access to a single automatically generated web page containing all the useful information regarding a specific event of interest. This is a change of paradigm from the retrieval of existing relevant content, to the generation of new documents as a synthesis of the relevant content [11]. Recent research works have studied new wikification algorithms to create dynamic Wikipedia pages, that are automatically edited based on social activity e.g. in Twitter [2]. Anyway, these methods do not consider any interaction with the end-user neither in the wikification nor in the consultation phase. Hence, users have no means to dynamically select or exclude some sources or to easily understand where some information is coming from. In addition, no feedback signals are gathered to improve wikification algorithms. For these reasons, this project is focused on the employment of visual analytics techniques to support the human-in-the-loop interaction and the comprehension of data provenance, which is exploited to improve both the wikification and event storylines generation processes.

Visual Analytics

Visual Analytics (VA) is "the science of analytical reasoning facilitated by interactive visual interfaces" [14]. VA integrates information visualization with data and model interaction with the purpose of helping the user to understand data and dynamically modify the algorithms underlying them. Progressive Visual Analytics (PVA) methods allow us to overcome the inefficiencies associated with the traditional "compute-wait-visualize" workflow. Besides, PVA methods allow analysts to inspect partial results of an algorithm without having to wait for the end of the process. The partial results of each stage are shown in the visual interface so that the user can make decisions that influence the progression of the analytical algorithms running in the background [6]. In Information Retrieval (IR) VA techniques have been applied recently to ease and make experimental evaluation more intuitive [4,5]. Whereas, despite being a promising and effective approach for dealing with streaming data, PVA has never been used in IR.

2 Project objectives

This project aims at developing innovative visual analytics tools to improve the wikification process and storyline generation, by exploiting dynamic and heterogeneous streaming data. Therefore, this project focuses on the employment of visual analytics techniques to support human-in-the-loop interactions and the use of data provenance to improve the quality of the wikification and event storylines generation processes and the user experience. In Figure 1, we see the user in the middle of the loop that generates articles. According to the human-in-the-loop model, there is a continuous interaction between the user and the visual interface. The system architecture we propose aims at maximizing the human contribution, which is fundamental to assess and improve the performances of the wikification and event storylines generation processes. Hence, we will focus on the application of VA methods, in a human-in-the-loop architecture in which the user feedback is useful to produce dynamic wiki articles that summarize the relevant information regarding a topic. This approach enhances data comprehension and exploits data provenance to reward the sources that provide more relevant and authoritative contents. This represents a change of paradigm, from static automatically generated articles to dynamic ones, in which the user becomes an active component for the improvement of wikification algorithms.

3 Research work description

The four main stages of this project are reported as follows:

1. **State-of-the-art inspection**: This stage focuses on studying the state-of-the-art of: web crawling, clustering algorithms for streaming data coming from social media, entity linking, wikification, data provenance, human-in-the-loop model and VA methods for dynamic interactive contents and data explainability.

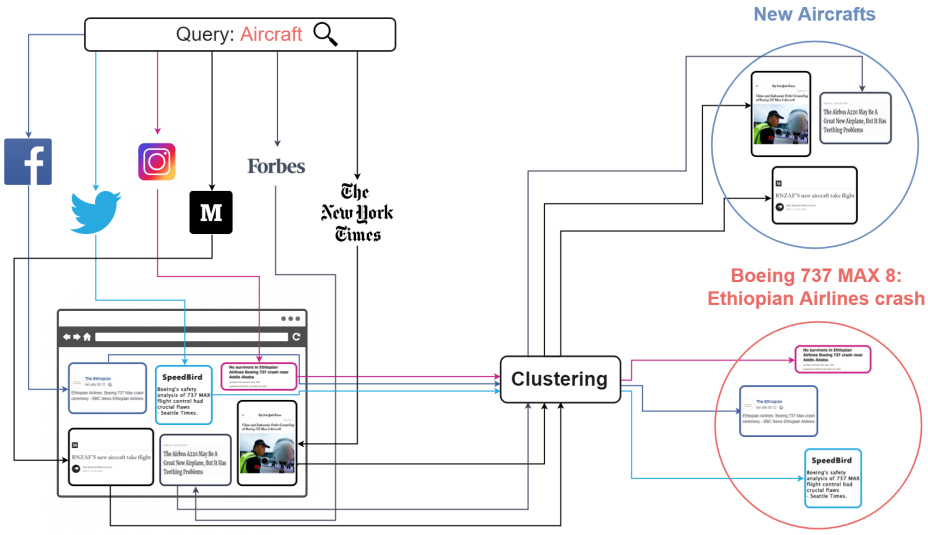

Fig. 2: Clustering of web crawled streaming data.

2. **Automatic wikification and event storylines generation**: This stage aims at reproducing state-of-the-art wikification algorithms and involves the following tasks (see the left part of Figure 1):
   - Web crawling and gathering of streaming data from social media, information networks and microblogging services.
   - Entity linking to the reference knowledge base. To this end, we will consider the use of relevant ontologies such as BabelNet[1].
   - Clustering of retrieved documents, articles, news and posts. Since streaming data come from heterogeneous multiple sources and information may be duplicated, clustering algorithms are necessary to aggregate semantically related documents. In Figure 2, we can see an example of the results for a not well specified query ("Aircraft"): the retrieved documents regard two different domains ("New Aircrafts" and "Ethiopian Airlines crash") and the purpose of clustering algorithms is to assign each document to the appropriate cluster. For clustering purposes, we exploit semantic information to enrich the bag-of-words (BOW) model and create a bag-of-concepts (BOC) document representation [8].
   - Event storylines reconstruction. Temporal information is exploited to produce timelines that present event happenings in chronological order. For this purpose, one possible benchmark dataset is presented in [16].

3. **Application of Visual Analytics (VA) methods**:
   This stage focuses on the application of VA methods to the wikified contents, generated in the previous phase. Therefore, during this stage, new VA tools for the automatic wikification will be developed. Since VA methods

[1] https://babelnet.org

rely on the human-in-the-loop model, according to which there is a continuous interaction between the user and the visual interface, VA interfaces play an important role. For this reason, this stage involves the study and development of intuitive VA interfaces designed to provide complete control of the parameters that influence the analytics algorithms designated for the extraction of useful information from data. To this aim, the choice of the UX framework (e.g. React[2]) for the development of the VA interfaces is crucial because interfaces need to be reactive and capable of updating quickly, according to the continuous flow of data coming from multiple streaming sources. In this context, the most relevant contents, selected by the analytical algorithms running in the background, are shown in the visual interface so that users can judge which contents are more relevant for the given topic to satisfy their information needs. The provided judgements act as useful feedback to improve the wikification algorithms and to allow the visualization of dynamic contents.

4. **Evaluation**: The last stage regards the evaluation of the overall architecture presented in Figure 1. In particular, this stage is focused on evaluating the performances of the wikification algorithms. The evaluation process will be done, by means of a tool that will be developed to compare different wikification algorithms, based on user assessments. Furthermore, different user studies will be done to investigate whether and how the developed tools improve the effectiveness of wikification algorithms and speed up access to knowledge. Some examples of user studies are: A/B testing, focus group, web analytics and first click testing.

4 Final remarks

This PhD project focuses on the application of VA methods to automatic wikification and event storylines generation. The employment of VA techniques allows us to generate dynamic wiki-like documents based on user feedback and interaction. In the last years, some research works have studied methods for storylines visualization [15], but the employment of visual analytics techniques for wikification and event storylines generation, still need to be examined. In addition, this project aims at investigating the combination of VA methods with algorithmic strategies, e.g. clustering of news and microblog posts [9]. It is worth noting that the automatic wikification and generation of event storylines are open problems. In recent years, plenty of work has been done to address these problems, but they are far to be solved. However, the efforts made to address these problems are certainly useful to improve the quality of automatically generated wiki documents and, most importantly, to ease the access to knowledge.

Acknowledgments: This work is partially supported by the Computational Data Citation (CDC-STARS) project of the University of Padua.

[2] https://reactjs.org

References

1. AlNoamany, Y., Weigle, M.C., Nelson, M.L.: Generating stories from archived collections. In: Proceedings of the 2017 ACM on Web Science Conference. pp. 309–318. WebSci '17, ACM, New York, NY, USA (2017)
2. Alonso, O., Kandylas, V., Tremblay, S.E.: Automatic story evolution wikification from social data. In: Twelfth International AAAI Conference on Web and Social Media (2018)
3. Alonso, O., Sellam, T.: Quantitative information extraction from social data. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 1005–1008. SIGIR '18, ACM, New York, NY, USA (2018)
4. Angelini, M., Fazzini, V., Ferro, N., Santucci, G., Silvello, G.: Claire: A combinatorial visual analytics system for information retrieval evaluation. Information Processing & Management **54**(6), 1077–1100 (2018)
5. Angelini, M., Ferro, N., Santucci, G., Silvello, G.: Virtue: A visual tool for information retrieval performance evaluation and failure analysis. Journal of Visual Languages & Computing **25**(4), 394–413 (2014)
6. Angelini, M., Santucci, G., Schumann, H., Schulz, H.J.: A review and characterization of progressive visual analytics. In: Informatics. vol. 5, p. 31. Multidisciplinary Digital Publishing Institute (2018)
7. Guo, B., Ouyang, Y., Zhang, C., Zhang, J., Yu, Z., Wu, D., Wang, Y.: Crowdstory: Fine-grained event storyline generation by fusion of multi-modal crowdsourced data. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **1**, 1–19 (2017)
8. Huang, A., Milne, D., Frank, E., Witten, I.H.: Clustering documents using a wikipedia-based concept representation. In: Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. pp. 628–636. PAKDD '09, Springer-Verlag (2009)
9. Li, L., Ye, J., Deng, F., Xiong, S., Zhong, L.: A comparison study of clustering algorithms for microblog posts. Cluster Computing **19**(3), 1333–1345 (2016)
10. Lin, C., Lin, C., Li, J., Wang, D., Chen, Y., Li, T.: Generating event storylines from microblogs. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. pp. 175–184. CIKM '12, ACM, New York, NY, USA (2012)
11. Lioma, C., Larsen, B., Petersen, C., Simonsen, J.G.: Deep learning relevance: Creating relevant information (as opposed to retrieving it) (2016)
12. Mihalcea, R., Csomai, A.: Wikify!: Linking documents to encyclopedic knowledge. In: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management. pp. 233–242. CIKM '07 (2007)
13. Rao, D., McNamee, P., Dredze, M.: Entity linking: Finding extracted entities in a knowledge base. In: Multi-source, Multilingual Information Extraction and Summarization. pp. 93–115. Springer Berlin Heidelberg (2013)
14. Scholtz, J.: Beyond usability: Evaluation aspects of visual analytic environments. In: 2006 IEEE Symposium On Visual Analytics Science And Technology. pp. 145–150. IEEE (2006)
15. Tanahashi, Y., Hsueh, C.H., Ma, K.L.: An efficient framework for generating storyline visualizations from streaming data. IEEE transactions on visualization and computer graphics **21**(6), 730–742 (2015)
16. Zubiaga, A.: A longitudinal assessment of the persistence of twitter datasets. Journal of the Association for Information Science and Technology **69**(8), 974–984 (2018)